14

Next-Generation Sequencing and Metagenomics

Nageswara Rao Reddy Neelapu[1] and Challa Surekha[2]

Abstract

Isolating microbes such as bacteria, fungi and virues from different environments like soil, water, air, plants, animals, hot springs, cold climates and space in purest form of culture is of interest for any biologist. Despite of many new technologies, limitations have stalled to get the purest forms of genetic material of microbes without losing its nativity. Microbes' losing their nativity means losing their ability to express and grow in the virtue of its unique environment. Metagenomics is a powerful tool to recover sample/the genetic material of a microbe or entire communities of organisms directly from its environment without losing its nativity.

This review summarizes information and current opinions in metagenomics by providing more insights on how gene expression has shaped the microbe development and also how an environment can influence the genome expression. How different next generation technologies had offered a platform to sequence the complete metagenomes of the microbes from the environment; and subjected to bioinformatics analysis such as sequence prefiltering, assembly, gene prediction, species diversity, data integration, and comparative metagenomics to deliver holistic information on interactions among communities, ecosystems and populations on community metabolism and population genomics. In addition what way, application of metagenomics has solved many challenges and proved its practicality in the fields of agriculture, food, engineering, evolution, medicine and sustainability.

1. Introduction

Metagenomics is the broad field with combination of microbiology, environment science and genomics commonly known as environmental genomics, ecogenomics or

[1.] Department of Biochemistry and Bioinformatics, School of Life Sciences, GITAM Institute of Science, GITAM University, Rushikonda, Visakhapatnam 530045 (AP), India Phone: +91-891-2840464 Fax: +91-891-2790032 Email: nrneelapu@gmail.com

[2.] Department of Biochemistry and Bioinformatics, School of Life Sciences, GITAM Institute of Science, GITAM University, Rushikonda, Visakhapatnam- 530045 (AP), India Phone: +91-891-2840464 Fax: +91-891-2790032 Email: challa_surekha@yahoo.co.in

[*] Corresponding author: nrneelapu@gitam.edu

community genomics. In nature when a sample was isolated and profiled, it revealed microbial communities which are rich and diverse. Traditional way of understanding an organism is by isolating it from the sample collected from the environment. Then cultivating the organism for pure clonal culture and sequencing the genome of the microbe to characterize its virtue. During this process microbes and their communities may lose their. Handelsman et al. (1998) for the first time used the term metagenomics to depict the idea of collection of genes sequenced directly from the environment. Chen and Patcher (2005) defined metagenomics as '...the application of modern genomics techinques to study the communities of microorganisms directly in their natural environments, by passing the need for isolation and lab cultivation of individual species'.

Metagenomics has its root in studies conducted on bacteria, eubacteria and archeabacteria based on rRNA sequences. These studies established the fact of existence of ribosome nucleotide sequence diversity along with microbial community diversity paving the way for present fields such as environmental genomics, ecogenomics or community genomics. Pace et al. (1991) proposed the idea of directly isolating and cloning bulk DNA from environmental sample. Healy et al. (1995) were successful in directly isolating and cloning bulk DNA from environmental sample consisting of microbial consortia by developing Zoo libraries. Stein et al. (1996) used marine samples to constructing marine libraries for sequencing of metagenome. Breitbart et al. (2002) used shot gun sequencing to establish the uncultured viral communities in 200 litres of sea water leading to the present day metagenomics. Subsequent studies cleared the twilight zone establishing new viruses in large numbers in human stool (Tyson et al. 2004), marine sediment (Tyson et al. 2004), and acid mine drainage system (Hugenhoz 2002). Developments in metagenomics have resulted in complete or nearly complete genome of bacteria/archeabacteria which were tough in obtaining pure cultures. Venter et al. (2004) as a part of Global Ocean Sampling Expedition (GOS) used shot gun metagenomics and indentified nearly 2000 different species, which include 148 types of bacteria never known before. Many prestigious projects are undertaken by different groups around the world. Among them Human Microbiome Initiative is one of the most prestigious project. Successful metagenomics project implements the following practices: sampling and processing, sequencing, sequence prefiltering, assembly, binning, annotation, sharing and storage of metadata and data analysis (Fig. 1.)

1.1 Sampling and processing

Sample processing is the crucial step in a metagenomics project. The DNA isolated shall be representative of all cells in the sample and with sufficient amounts of high-quality nucleic acids. Sample is associated with a host (e.g. an animal or plant), then either fractionation or selective lysis is used (Burke et al. 2009; Thomas et al. 2010). Physical fractionation can be used if community sample is viruses in seawater. Direct lysis of the soil sample is quantifiable than indirect lysis. Biopsies or groundwater yield very small amounts of DNA. Thus, sampling and processing is an important step for getting required quantities of DNA for sequencing.

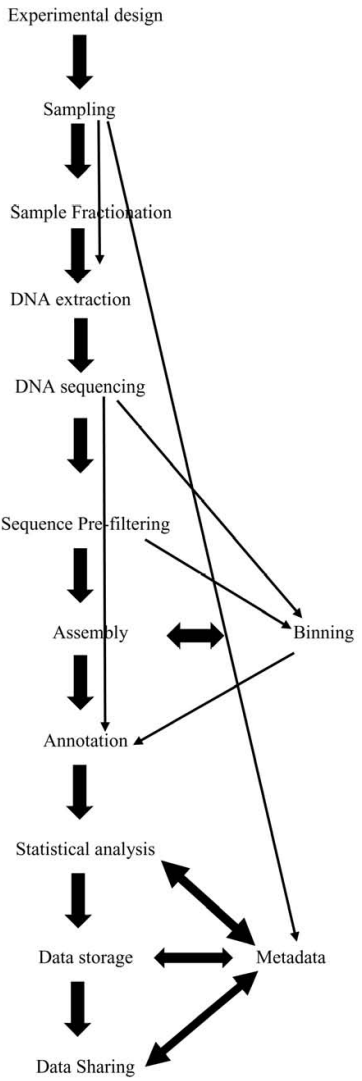

FIGURE 1. Flow diagram of metagenome project

1.2 Metagenome sequencing technology

DNA double stranded structure was established in the year 1953 (Watson and Crick 1953), but it took decades to develop DNA sequencing methods. Ray Wu (1970) developed the first method to determine the DNA sequence using location specific primer extension strategy. Synthetic location – specific primer was used to determine the sequence between the years 1970 and 1973 (Padmanabhan and Wu 1972; Wu et al. 1973; Jay et al. 1974) and Wu and Taylor (1971) subjected phage λ DNA to sequence 12 bp of cohesive ends. Chemical degradation was used by Walter Gilbert and Allan Maxam to sequence DNA (Maxam and Gilbert 1977). More rapid DNA

sequencing method with chain-terminating inhibitors was developed by Frederick Sanger, which is now considered as the first generation sequencing method (Sanger et al. 1977). This method identifies linear randomly terminated nucleotide sequences, whereas automated Sanger's method uses fluorescent labeled terminators. The most accurate, most available and well defined technology is Sanger sequencing method. The throughput of this automated method can read 96 reactions in parallel. Using this method, complete human genome would take around 60 years to synthesize and cost approximately $5 to 30 million USD. First semi-automated DNA sequencing machine was announced in 1986 by Leroy E. Hood's laboratory followed by fully automated sequences by Applied Biosystem in 1987 and novel florescent labeling technique by Dupont's Genesis 2000 (Prober et al. 1987). Next generation high throughput DNA sequencing technologies capable of sequencing large number DNA sequences in a single reaction were developed. Next generation sequencing (NGS) technology provide high speed, throughput and monitors the sequential addition of nucleotides to DNA templates. The need and demand for low cost sequencing made the development of next generation sequencing (high-throughput sequencing) technologies. Overall high cost, reduction of sequencing errors, low reading accuracy are the limiting factors of the new technologies.

Several NGS methods like 454 pyrosequencing, llumina (Solexa) sequencing, SOLiD Sequencing, Massively parallel signature sequencing (MPSS), Polony sequencing, Ion Torrent semiconductor sequencing, DNA nanoball sequencing, Heliscope single molecule sequencing and Single molecule real time (SMRT) sequencing etc are the technologies used for sequencing metagenomes. Genomic DNA, cDNA, immunoprecipitated DNA can serve as DNA template for all NGS experiments (Fig. 2)

TABLE 1. Highlights of next generation sequencing technologies

Single-molecule real time sequencing (SMRT)	
Read Length	10,000bp to 15,000bp avg (14,000 bp N50); maximum read length >40,000 bases
Accuracy	9999% consensus accuracy; 87% sigle read accuracy
Reads per run	50.000 per SMRT cell or 500-1000 megabases
Time per run	30 minutes to 4 hours
Cost per 1 million bases	$0.13-0.60
Advantages	Longest read length, Fast detects 4mC, 5mC, 6mA.
Disadvantages	Moderate throughput Equipment can be very expensive
Ion Semiconductor (Ion Torrent sequencing)	
Read Length	Upto 400bp
Accuracy	98%
Reads per run	Upto 80 million
Time per run	2 hours
Cost per 1 million bases	$1
Advantages	Less expensive equipment, Fast
Disadvantages	Hompolymer errors
Pyrosequencing (454)	
Read Length	750bp
Accuracy	99.9%
Reads per run	1million
Time per run	24hours

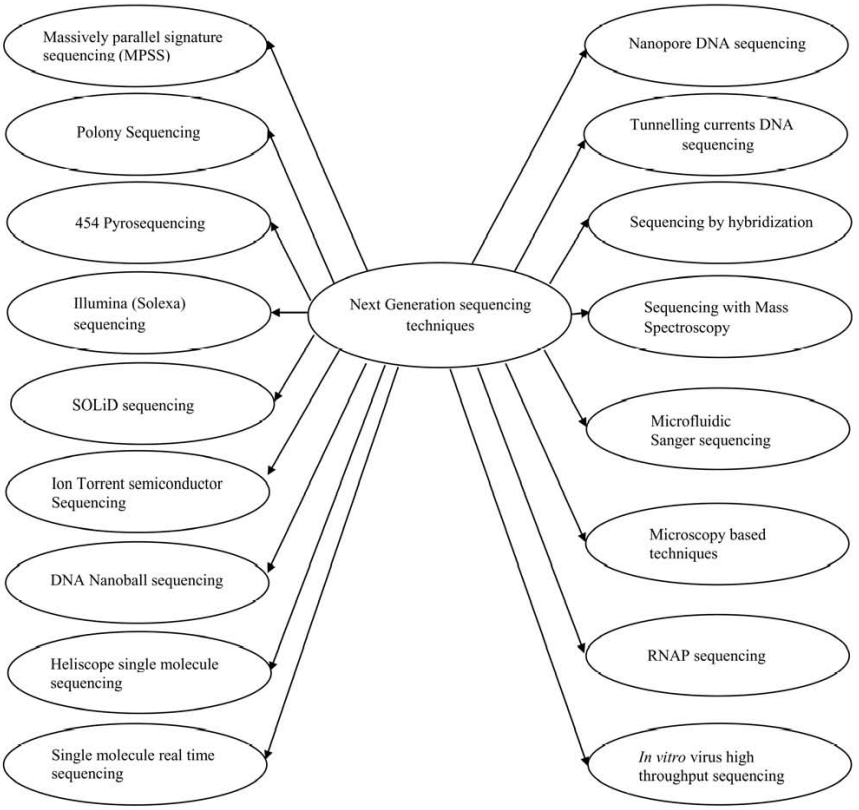

FIGURE 2. Next generation sequencing technologies

Cost per 1 million bases	$10
Advantages	Long read size, Fast
Disadvantages	Runs are expensive, Homopolymer errors
Solexa Sequencing synthesis (Illumina)	
Read Length	500 to 300 bp
Accuracy	98%
Reads per run	Upto 3 million
Time per run	1 to 10days depending on sequenceand specified read length
Cost per 1 million bases	$0.05 to 0.12
Advantages	Potential for high sequence yield, depending upon sequencer model and desired application
Disadvantages	Equipment can be very expensive and Requires high concentrations of DNA.
Sequencing by Ligation (SOLiD sequencing)	
Read Length	50+35 or 50+ 50 bp
Accuracy	99%
Reads per run	1.2 to 1.4 billion
Time per run	1 to 2 weeks

Cost per 1 million bases	$0.13
Advantages	Low cost per base
Disadvantages	Slowerr than other methods, have issue sequencing palindromic sequence
Chain termination (Sanger sequencing)	
Read Length	400 to 900 bp
Accuracy	99.9%
Reads per run	N/A
Time per run	20 minutes to 3 hours
Cost per 1 million bases	$2400
Advantages	Long individual reads, Useful for many applications
Disadvantages	More expensive and impractical for larger sequencing projects.

1.2.1 454 pyrosequencing

454 pyrosequencing was first reported in 1988 using the principle of pyrophosphate detection and the technique was further developed (Ronaghi et al. 1996) in Stockholm to analyze 96 samples in parallel in a microtiter plate. The 454 GenomeSequencer FLX instrument was introduced in 2005 by 454 Life Sciences. Pyrosequencing method is based on principle 'sequencing by synthesis'. Template DNA is ligated with specific adapter to attach one fragment of DNA to a single primer coated on streptavidin bead to form a clonal colony. Emulsion PCR is carried out with water droplets in an oil solution. The droplets act as individual amplification reactors producing approximately 10^7 clonal copies per bead (Margulies et al. 2005). ATP sulfurylase and luciferase enzymes generate visible light upon incorporation of the complementary nucleotide. The amplified DNA is transferred into a picotiter plate and analyzed using a pyrosequencing reaction. The picotiter plate allows hundreds and thousands of parallel pyrosequencing reactions to be carried out increasing the sequencing throughput. As this approach is sequencing-by-synthesis, it measures the release of inorganic pyrophosphate detected by light enzyme luciferase (Tawfik and Griffiths 1998; Nyren et al. 1993). The light intensity is proportional to the pyrophosphate released. This technology provides intermediate read length, generates 80-120Mb of sequence in 4 hours.

1.2.2 Illumina (Solexa) sequencing

Illumina (Solexa) sequencing is based on reversible dye-terminators technology, and engineered polymerases (Bentley et al. 2008). This approach was invented by Balasubramanian and Klennerman from Cambridge University's chemistry department. This method is similar to Sanger sequencing, but uses modified dNTP's containing a terminator which blocks further polymerization. DNA molecules (single stranded) are attached to an immobilized surface by an adapter, subsequently bent and hybridized to complementary adapters for synthesis of their complementary strands. After the amplification, the templates are sequenced in a massive parallel form using sequencing-by-synthesis approach. To determine the sequence, four types of reversible terminator bases (RT-bases) are added in presence of special DNA polymerases and non-incorporated nucleotides are washed away (Ronaghi et al. 1996). A camera takes images of the fluorescently labelled terminator nucleotides and its position. Then the dye along with the terminal 3' blocker, are chemically removed from

the DNA, allowing for the next cycle to begin. The Illumina sequencing is capable of generating 35bp reads produce 1Gb sequence in 2-3 days. The method is highly accurate base by base sequencing by eliminating errors.

1.2.3 SOLiD sequencing

SOLiD sequencing employs the process by ligation and was introduced in autumn 2007. A library of DNA fragments are first ligated to magnetic bead with universal P1 adapter. Emulsion PCR takes place in microreactors with all reagents of PCR. The resulting beads, each containing single copies of the same DNA molecule, are deposited on a glass slide (Valouev et al. 2008). One DNA fragment per bead bound to an adapter is hybridized with a primer. These templates are characterized by fluorescent labels and detected by fluorescence. The sequencing process is continued in the same way with another primer and sequence reading length is about 35 bases. Sequences are determined in parallel for more than 50 million bead clusters, resulting in a very high throughput of the order of Gigabases per run. The new SOLiD instrument is capable of producing 1-3Gb of sequence in 8-day run and offers 99.94% accuracy.

1.2.4 Massively parallel signature sequencing (MPSS)

The first of the next-generation sequencing technologies, MPSS, was developed in the 1990s at Lynx Therapeutics by Sydney Brenner and Sam Eletr. MPSS is a procedure used for identifying and quantifying mRNA transcripts similar to serial analysis of gene expression (SAGE). MPSS was a bead-based method that used a complex approach of adapter ligation followed by adapter decoding, reading the sequence in increments of four nucleotides. mRNA is reverse transcribed into cDNA, cDNA is amplified in microreactors as emulsion PCR. A sequence signature of ~16-20bp is determined from all the beads in parallel. Each signature sequence is cloned onto microbeads and then arrayed in a flow cell for sequencing and quantification. This method made it susceptible to sequence-specific bias or loss of specific sequences. The technology was so complex, MPSS was only performed 'in-house' by Lynx Therapeutics and no DNA sequencing machines were sold to independent laboratories. Lynx Therapeutics merged with Solexa (later acquired by Illumina) in 2004, leading to the development of sequencing-by-synthesis, a simpler approach acquired from Manteia Predictive Medicine, which rendered MPSS obsolete. Typically used for sequencing cDNA to measure gene expression levels (Brenner et al. 2000).

1.2.5 Polony sequencing

Polony sequencing was developed in the laboratory of George M Church at Harvard. This technique was among the first next-generation sequencing systems and was used to sequence a full genome in 2005. It combined an *in vitro* paired-tag library with emulsion PCR, an automated microscope, and ligation-based sequencing . The chemistry to sequence an *E. coli* genome at an accuracy of >99.9999% costed approximately 1/9 that of Sanger sequencing (Shendure et al. 2005). Emulsion PCR isolates individual DNA molecules along with primer-coated beads in aqueous droplets within an oil phase. PCR then coats each bead with clonal copies of the DNA molecule followed by immobilization for later sequencing.

1.2.6 Ion Torrent semiconductor sequencing

Ion Torrent Systems Inc. developed a system based on using standard sequencing chemistry, but with a novel, semiconductor based detection system. This method of sequencing is based on the detection of hydrogen ions that are released during the polymerisation of DNA, as opposed to the optical methods used in other sequencing systems. A microwell containing a template DNA strand to be sequenced is flooded with a single type of nucleotide. If the introduced nucleotide is complementary to the leading template nucleotide it is incorporated into the growing complementary strand. This causes the release of a hydrogen ion that triggers a hypersensitive ion sensor, which indicates that a reaction has occurred. If homopolymer repeats are present in the template sequence multiple nucleotides will be incorporated in a single cycle. This leads to a corresponding number of released hydrogens and a proportionally higher electronic signal (Rusk 2011).

1.2.7 DNA nanoball sequencing

DNA nanoball sequencing is a type of high throughput sequencing technology used to determine the entire genomic sequence of an organism. The method uses rolling circle replication to amplify small fragments of genomic DNA into DNA nanoballs. Fluorescent probes bound to complementary DNA are ligated to anchor sequences bound to known sequences on the DNA template. Unchained sequencing by ligation is then used to determine the nucleotide sequence (Drmanac et al. 2010). This method of DNA sequencing allows large numbers of DNA nanoballs to be sequenced per run and at low reagent costs compared to other next generation sequencing platforms (Porreca 2010). However, only short sequences of DNA are determined from each DNA nanoball which makes mapping the short reads to a reference genome difficult. This technology has been used for multiple genome sequencing projects and is scheduled to be used for more.

1.2.8 Heliscope single molecule sequencing

Heliscope sequencing is a method of single-molecule sequencing developed by Helicos Biosciences. This method helps in direct sequencing of cellular and extracellular nucleic acids in an unbiased manner. It uses DNA fragments with added poly-A tail adapters which are attached to the flow cell surface. The next steps involve extension-based sequencing with cyclic washes of the flow cell with fluorescently labeled nucleotides (one nucleotide type at a time, as with the Sanger method). The reads are performed by the Heliscope sequencer. The reads are short, up to 55 bases per run, but recent improvements allow for more accurate reads of stretches of one type of nucleotides (Heliscope gene sequencing; Thompson and Steinmann 2010). This sequencing method and equipment were used to sequence the genome of the M13 bacteriophage (Harris et al. 2008).

1.2.9 Single molecule real time (SMRT) sequencing

SMRT sequencing is based on sequencing by synthesis approach. The DNA is synthesized in zero-mode wave-guides (ZMWs) – small well-like containers with the

capturing tools located at the bottom of the well. The sequencing is performed with use of unmodified polymerase (attached to the ZMW bottom) and fluorescently labelled nucleotides flowing freely in the solution. The wells are constructed in a way that only the fluorescence occurring by the bottom of the well is detected. The fluorescent label is detached from the nucleotide at its incorporation into the DNA strand, leaving an unmodified DNA strand. According to Pacific Biosciences, the SMRT technology developer, this methodology allows detection of nucleotide modifications (such as cytosine methylation). This happens through the observation of polymerase kinetics. This approach allows reads of 20,000 nucleotides or more, with average read lengths of 5 kilobases.

1.2.10 New Novel sequencing methods in development

There are a number of novel DNA sequencing methods which are still in development. The development is made in terms of reduction of reaction volumes, smaller amounts of reagents and low cost. Third generation technologies are aiming to increase the throughput, decrease time and cost, harnessing the processivity of DNA polymerase (Schadt et al. 2010). Methods in development are Nanopore DNA sequencing, Tunneling currents DNA sequencing, Sequencing by hybridization, Sequencing with mass spectrometry, Microfluids Sangers sequencing, Microscopy based technique, RNAP sequencing, *Invitro* virus high-throughput sequencing.

1.2.11 Applications of Next generation sequencing

Apart from metagenomics NGS can be applied to genome sequencing & resequencing, transcriptome profiling, DNA – protein interactions and epigenome characterization and resurrection of ancient genome. (Fig 3).

Transcriptome sequencing

Genome wide survey of gene expression levels were studied using qPCR, SAGE and microarray with limitations. Next generation sequencing techniques were implemented along with SAGE tags to sequence the RNA populations from the cells expressed. Noncoding RNA (ncRNA) are any RNA's that are transcribed and not translated to a protein. ncRNA in plants and animals are having an important role in regulation of gene expression. Next-generation sequencing technology has discovered many novel ncRNAs (Sanger and Coulson 1975; Venter et al. 2004; Nyren et al. 1993; Ewing 1998; Bainbridge et al. 2006; Hutchison 2007; Gowda et al. 2006; Mardis, 2006). These ncRNA are unique, diverse and regulate genes by a variety of mechanisms. The readouts from next generation technologies are quantitative, allowing to detect changes in expression levels due to changes in environment and onset of disease. Studying the roles of these new specific RNAs may help in uncover certain aspects of disease or cancer. Remarkable progress has been made in characterizing and understanding these molecules using next generation technologies. Discovering ncRNAs and sequencing of transcriptome would provide us new insights on the genome wide expression patterns of an organism.

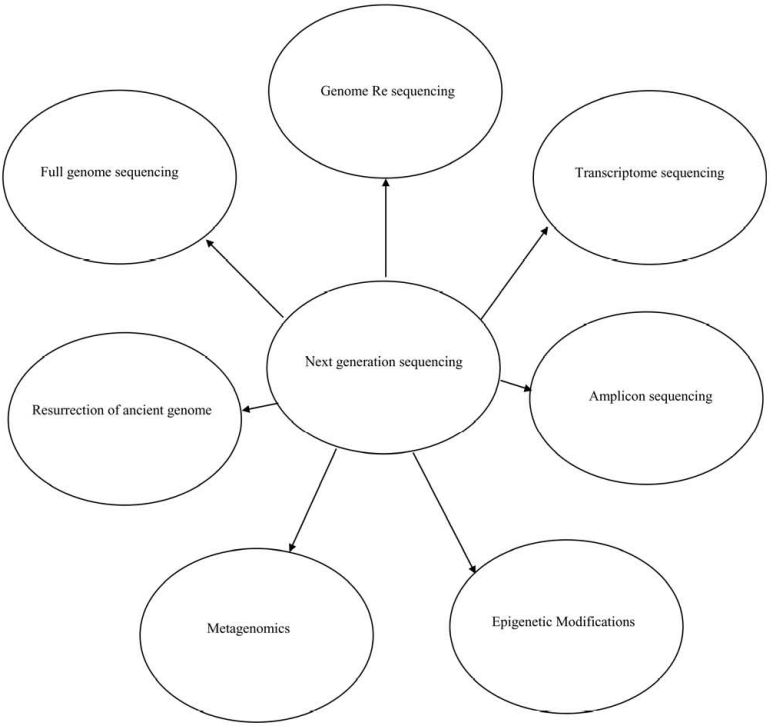

FIGURE 3. Applications of next generation technologies

Resurrection of ancient genomes

Samples from fossils and ancient remnants are in a state of degradation. With the advent of molecular techniques such as PCR and DNA sequencing, deciphering of mitochondrial genomes from fossils and ancient remnants was possible. Several non-trivial technical complications arised from these cases, most notably DNA contamination. Next generation sequencing can be implemented to obtain sequence information from the degraded nature of the ancient genome. Sequence information from single fossil bone of Neanderthal genome was obtained using next generation sequencing technologies (Carrilho 2000). In addition to Neanderthal sequence, information was also obtained directly from the nuclear genomes of ancient remains of the cave bear (Robertson et al. 2007) and mammoth (Barski et al. 2007). In this way next generation sequencing technology can be used in resurrecting ancient genomes.

Analysis of epigenetic modifications of histones and DNA

Next generation sequencing technologies made it possible to study DNA methylation profile by bisulfite DNA sequencing, mapping histone modifications, mapping the locations of DNA-binding proteins, DNA accessibility and chromatin structure. The association between DNA and proteins is an interaction regulating gene expression and controlling the availability of DNA for transcription, replication, and other processes. Genome-wide chromatin immunoprecipitation (ChIP)-based studies for

DNA-protein interactions became possible in sequenced genomes (Korbel et al. 2007). ChIP-based approach and the Illumina platform provided insights into transcription factor binding sites in the human genome such as neuron-restrictive silencer factor (NRSF) (Tawfik and Griffiths 1998) and signal transducer and activator of transcription 1 (STAT1) (Bhinge et al. 2007). So, next generation sequencing technologies can be used for understanding of gene expression–based cellular responses due to DNA-protein interactions.

1.3 Sequence prefiltering

After obtaining metagenomic data, prefiltering of the sequence is the first step. Prefiltering includes removal of redundant, low quality sequences and sequences of eukaryotic origin (Adey et al. 2010; Bentley et al. 2010) using the methods EuDetect and DeConseq (Nakamura et al. 2011; Hess et al. 2011).

1.4 Assembly

Recovering the sequence from the environmental sample is the first important step followed by assembly of the recovered reads into metagenome. Two strategies are used for assembly of metagenomics samples: de novo assembly and reference-based assembly (co-assembly). De novo assembly tools are developed based on the de Bruijn graphs to compute large amounts of data. de Bruijn assemblers Velvet (Zerbino and Birney 2008) or SOAP (Li et al. 2008) are the tools used for De novo assembly. Reference based assembly can be used, if closely related reference genomes are available for assembly of the metagenomic dataset. The available reference based assembly software are Newbler (Roche), AMOS, or MIRA (Chevreux et al. 1999). Using the appropriate method for assembly of the metagenomic dataset is important.

1.5 Binning

Binning is….. 'the process of sorting DNA sequences into groups that might represent an individual genome or genomes from closely related organisms……'. Binning algorithms employ two types of information contained within a given DNA sequence. Firstly, compositional binning uses conserved nucleotide composition and secondly, the similarity of with a reference database can be used to classify and bin the sequence. Phylopythia (McHardy et al. 2007), S-GSOM (Chan et al. 2008), PCAHIER (Diaz et al. 2009; Zheng and Wu 2010) and TACAO (Diaz et al. 2009) are the compositional-based binning algorithms, whereas IMG/M (Markowitz et al. 2008), MG-RAST (Glass et al. 2010), MEGAN (Huson et al. 2007), CARMA (Krause et al. 2008), SOrt-ITEMS (Haque et al. 2009) and MetaPhyler (Liu et al. 2009) are similarity-based binning softwares. PhymmBL (Brady and Salzberg 2009) and MetaCluster (Leung et al. 2011) are the binning algorithms that consider both composition and similarity.

1.6 Annotation

Annotation of metagenomes can be performed by two different initial pathways. RAST or IMG the existing pipelines for genome annotation are used if the genome is

reconstructed or annotation on the entire community. Later metagenomic sequence data can be annotated in two steps: first, genes are identified and predicted; and second, putative gene functions are assigned. CDS of metagenome are predicted using tolls such as FragGeneScan (Rho et al. 2010), MetaGeneMark (McHardy et al. 2007), MetaGeneAnnotator (MGA)/ Metagene (Noguchi et al. 2008) and Orphelia (Hoff et al. 2009; Yok and Rosen 2011). BLAST-based searches are used for functional annotation. KEGG (Kanehisa et al. 2004), eggNOG (Muller et al. 2010), COG/KOG (Tatusov et al. 2003), PFAM (Finn et al. 2010), and TIGRFAM (Selengut et al. 2007) are the reference databases giving functional context to metagenomic datasets.

1.7 Sharing and storage of metadata

Genomic research has witnessed and is following the tradition of sharing genomic data as public databases. But, metagenomics field is yet to witness this tradition of sharing metagenomic data. At present NCBI, IMG/M, CAMERA and MG-RAST are the new level of organization and collaboration that provide metadata and centralized services.

1.8 Application of metagenomics

Metagenomics has been applied to solve many challenges in the fields of medicine, biofuel, enviromental remediation, biotechnology, agriculture and ecology (Fig 4.)

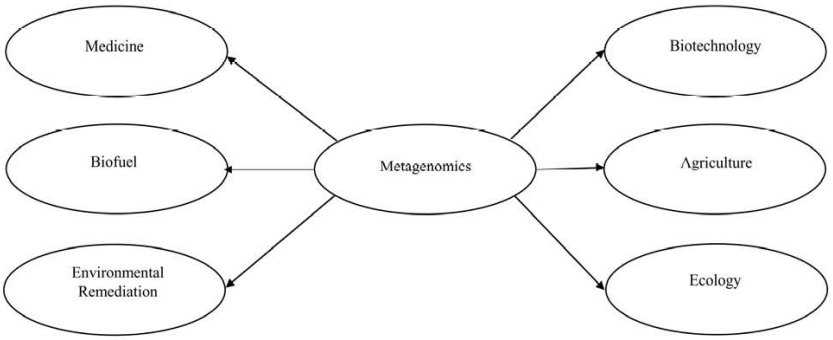

FIGURE 4 Applications of metagenomics

1.8.1. Medicine

Two metagenomic projects Human Microbiome initiative and Metagenomics of the Human Intestinal Tract (MetaHit) established the fact that microbial communities play an important role in human health. Human Microbiome initiative characterized microbial communities from 15-18 body sites of 250 individuals. The primary goal of this project is to correlate the human microbiome with human health. This project was somewhat successful in understanding the key role of microbial communities in preserving human health. Some more questions like composition of microbial

TABLE 2. Metagenome assemblers used for assembling reads into metagenomes

S. No	Assembler	Description	Reference
1	Orione	A Galaxy-based framework consisting of publicly available research software and specifically designed pipelines to build complex, reproducible workflows for next-generation sequencing microbiology data analysis.	Cuccuru et al., 2014
2	GARM	A new software pipeline to merge and reconcile assemblies from different algorithms or sequencing technologies.	Soto-Jimenez et al., 2014
3	GeneStitch	Novel way of using the de Bruijn graph assembly of metagenomes to improve the assembly of genes.	Wu et al., 2012
4	CLC Bio's denovo assembler (CLC Bio)	CLC bio's de novo assembly algorithm utilize de Bruijn graphs to represent overlapping reads which is a common approach for short read de novo assembly	http://www.clcbio.com/files/appnotes/CLC_bio_De_novo_Assembly.pdf
5	Ray Meta	A massively distributed metagenome assembler that is coupled with Ray Communities, which profiles microbiomes based on uniquely-colored k-mers	Boisvert et al., 2012
6	MAP	A de novo metagenomic assembly program for shotgun DNA reads	Lai et al., 2012
7	MetAMOS	A modular and open source metagenomic assembly and analysis pipeline	Treangen et al., 2013
8	Meta-IDBA	An iterative De Bruijn Graph De Novo short read assembler specially designed for de novo metagenomic assembly	Peng et al., 2007
9	MetaVelvet	We modified and extended a single-genome and de Bruijn-graph based assembler, Velvet, for de novo metagenome assembly	Namiki et al., 2012
10	Newbler	Newbler is for de novo DNA sequence assembly. It is designed specifically for assembling sequence data generated by the 454 GS-series of pyrosequencing platforms sold by 454 Life Sciences, a Roche Diagnostics company.	Nederbragt, 2014
11	MIRA	Assembler uses iterative multipass strategies centered on high-confidence regions within sequences and has a fallback strategy for using low-confidence regions when needed	Chevreux et al., 2004
12	SPADES	Assembler for both single-cell and standard (multicell) assembly	Bankevich et al., 2012

communities, core human microbiome etc are yet to be answered (Kristensen et al. 2009; Zimmer 2010).

The MetaHit project is based on 124 individuals related to healthy and diseased states like overweight and irritable bowel syndrome. The study established bacteriocides and firmicutes as the dominant distal gut bacteria and identified 1244 metagenomic clusters important for the health of the intestinal tract. These clusters were

of two types : housekeeping and intestine specific. The housekeeping bacteria are required to influence the central metabolic pathways of the host like central carbon metabolism and aminoacid synthesis. The intestine/gut-specific bacteria functions include adhesion to host proteins or harvesting sugars of globoseries and glycolipids. Other outcomes of the project is that change in gut biome diversity was observed in patients with irritable bowel syndrome when compared with healthy individuals. Another new insight provided by this project was that only 7.6% of the known gut bacteria were captured and more research is necessary to identify novel bacteria from the gut biome.

1.8.2. Biofuel

Biomass such as corn and its stalk, grasses, sugarcane etc are used for the conversion of the cellulose into fuels which are known as biofuel. Different types of biofuels produced include ethanol, methane and hydrogen. Conversion of biomass into biofuel requires microbial consortia. This consortium converts/transforms various complex carbohydrates like cellulose into simple sugars. Later these sugars are fermented to produce ethanol, sometimes methane and hydrogen as main/by products.

Metagenomics is a poweful tool which helps us in understanding and providing new insights on how these microbial communities achieve the required function in the particular environment. This information can later be applied to control microbial communities for achieving the required function in controlled environment. Luen-Luen et al. (2009) applied metagenomics to screen enzymes like glycoside hydrolases from microbial consortia involved in biofuel production. Jaenicke et al. (2011) applied metagenomics to understand microbial consortia of biogas fermentors. Suen et al. (2010) applied metagenomics to understand the role of gut microbiome of insects-leaf cutter ants in converting biomass to simple substances in the gut of insect. So, metagenomics has helped in deciphering the role of microbial communities in converting complex biomass into simple molecules. Thus, this reinvented technology can be implemented in production of biofuels.

1.8.3. Enviromental remediation

Monitoring the impacts of pollutants on the environment and cleaning up the pollutants from the environment is the biggest task today. The amount of pollutants released in the environment is increasing leading to presence of huge amounts of xenobiotics. Some are easily degraded by the enzymes and the rest of amount is accumulated in the soil, water and living organisms. These accumulated pollutants are toxic and cause mutations finally leading to cancer in the organisms. Interestingly some of the organisms are resistant, rendering pollutants non-toxic and making the microbes adapt to polluted environment. The proteins, enzymes and genes of microbes had a capacity to break down the pollutants. Bioremediation can be *exsitu* by removing pollutants and *insitu*. Microbes remediating pollutants help to have an eco-friendly and cost effective way to have a health in polluted ecosystems. Metagenomics can be applied to isolate, assess and understand how microbial communities are coping up with pollutants and assess the improvement in contaminated sites (George et al. 2010).

1.8.4. Biotechnology

A vast array of biologically active compounds were identified and recovered from microbes. These compounds are actively being used as fine chemicals, agrochemicals and pharmaceuticals. Application of metagenomics helps in identification of desired trait or useful activity (Wong 2010). There are two types of bioprospecting metagenomic data: function-driven screening of expression data and sequence-driven screening of DNA sequence data (Patrick and Handelsman 2003). Function-driven screening analysis identifies clones with desirable traits or useful activity followed by biochemical characterization and sequence analysis. Whereas sequence-driven analysis uses PCR primers to screen clones for sequence of interest (Patrick and Handelsman 2003). Sometimes a combination of function and sequence-driven screening can be used for bioprospecting.

1.8.5. Agriculture

Plants are surrounded by a number of microbial communities both on it and in the soil. Much is not known about the microbial consortia inhabitating the soil despite of their economic importance. Microbial communities are known to fix atmospheric nitrogen, inducing plant growth, nutrient recycling, sequestering of iron and other metals and disease supression. Metagenomics can be used for exploring the microbial communties interacting with plants to improve crop health.

1.8.6. Ecology

Environmental communities is the well-known application for metagenomics. Community genomics can be used to understand the role of each community in an ecosystem. There are two types of microbial communities identified from metagenomic analysis : feast/famine and planktons. Metagenomics can provide more information and insights in functions of microbial communities (Raes et al. 2011). Meta genomics was applied to identify microbial consortia found in faeces of Australian sea lions. Australain sea lions faeces are rich with nutrients and microbial consortia. Microbial consortia present in the faeces break down the nutrients in the faeces and make it available to the food chain of the coastal ecosystem (Lavery et al. 2012). In addition, metagenomics also helps us in identifying microbial communities present in air, water and debris.

1.8.7. Metagenomics for species/strain Identification

Efforts to unambiguously classify metagenomic reads into species/strains/higher levels/clade-specific is challenging. Next generation technologies are used to capture DNA/RNA from the samples. These metagenome reads captured in the form of DNA/RNA are used to perform computationaly extensive assembly to identify species/strains. Challenges in this process are sequencing errors, noise generated in reads during sequencing, ambiguity contributed by homology in the genome content of closely related species/strains, complex data preprocessing and assembely based on genome coding regions (Wang et al. 2012). Tu et al. (2014) considering both genome coding and non-coding data developed K- mer approach to identify species/

strains from a metagenome data. *K*-mer-based approach, identifies genome-specific markers (GSMs) rapidly and comprehensively from all regions in the genome sequence and filter out non-specific sequences. The currently sequenced microbial metagenomes are searched against these GSMs, to determine the presence/absence and/or the relative abundance of each strain/species. Thus, species/strain identification is possible.

Acknowledgements

We would like to thank GITAM University, Visakhapatnam, India for providing the facility and support. The authors also thankful to Prof. I. Bhaskar Reddy and Dr Malla Rama Rao for constant support throughout the research work.

References

Ray, W. *Faculty Profile*. Cornell University. Archived from the original on 2009-03-04.

Adey, A., H.G. Morrison, X. Asan Xun, J.O. Kitzman, and E.H. Turner. 2010. Rapid, low-input, lowbias construction of shotgun fragment libraries by high-density *invitro* transposition. *Genome Biol*. 11(12): R119.

Bainbridge, M.N., R.L. Warren, M. Hirst, T. Romanuik, T. Zeng, A. Go, A. Delaney, M. Griffith, M. Hickenbotham, V. Magrini, E.R. Mardis, M.D. Sadar, A.S. Siddiqui, M.A. Marra, and S.J. Jones. 2006. Analysis of the prostate cancer cell line LNCaP transcriptome using a sequencing-by-synthesis approach. *BMC Genomics* 7 : 246.

Bankevich, A., S. Nurk, D. Antipov, A.A. Gurevich, M. Dvorkin, A.S. Kulikov, V.M. Lesin, S.I. Nikolenko, S. Pham, A.D. Prjibelski, A.V. Pyshkin, A.V. Sirotkin, N. Vyahhi, G. Tesler, M.A. Alekseyev, and P.A. Pevzner. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol.* 19 (5) : 455-77.

Barski, A., S. Cuddapah, K. Cui. T.Y. Roh, D.E. Schones, Z. Wang, G. Wei, I. Chepelev, and K. Zhao. 2007. High-resolution profiling of histone methylations in the human genome. *Cell.* 129 : 823–37.

Bentley, D.R., S. Balasubramanian, H.P. Swerdlow, G.P. Smith, and J. Milton, et al. 2008. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 456 (7218): 53-59.

Bhinge, A.A., J. Ki, G.M. Euskirche, M. Snyder, and V.R. Iyer. 2007. Mapping the chromosomal targets of STAT1 by sequence tag analysis of genomic enrichment (STAGE). *Genome Res*. 17 : 910–916.

Boisvert, S., F. Raymond, E. Godzaridis, F. Laviolette, and J. Corbeil. 2012. Ray Meta: scalable de novo metagenome assembly and profiling. *Genome Biol*. 13 (12) : R122.

Brady, A., and S.L. Salzberg. 2009. Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models. *Nat Methods*. 6 (9): 673-676.

Breitbart, M., P. Salamon, B. Andresen, J.M. Mahaffy, A.M. Segall, D. Mead, F. Azam, and F. Rohwer. 2002. Genomic analysis of uncultured marine viral communities. *Proc. Natl. Acad. Sci. U.S.A*. 99 (22): 14250–255.

Brenner, S., M. Johnson, J. Bridgham, G. Golda, D.H. Lloyd, D. Johnson, S. Luo, S. McCurdy, M. Foy, M. Ewan, R. Roth, D. George, S. Eletr, G. Albrecht, E. Vermaas, S.R. Williams, K. Moon, T. Burcham, M. Pallas, R.B. DuBridge, J. Kirchner, K.

Fearon, J. Mao, and K. Corcoran. 2000. Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nat Biotechnol*. 18(6):630-34.

Burke, C., S. Kjelleberg, and T. Thomas. 2009. Selective extraction of bacterial DNA from the surfaces of macroalgae. *Appl Environ Microbiol*. 75(1):252-56.

Carrilho, E. 2000. DNA sequencing by capillary array electrophoresis and microfabricated array systems. *Electrophoresis* 21: 55–65.

Chan, C.K., A.L. Hsu, S.K. Halgamuge, S.L. Tang. 2008. Binning sequences using very sparse labels within a metagenome. *BMC Bioinformatics*. 9: 215.

Chen, K., and L. Pachter. 2005. Bioinformatics for whole-genome shotgun sequencing of microbial communities. *PLoS Comput. Biol*. 1 (2): e24.

Chevreux, B., T. Pfisterer, B. Drescher, A.J. Driesel, W.E. Müller, T. Wetter, and S. Suhai. 2004. Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. *Genome Res.* 14: 1147-1159.

Chevreux, B., T. Wetter, and S. Suhai. 1999. Genome sequence assembly using trace signals and additional sequence information computer science and biology. Proceedings of the German Conference on *Bioinformatics*. 99: 45-56.

CLC_bio_De_novo_Assembly. http://www.clcbio.com/files/appnotes/CLC_bio_De_novo_Assembly.pdf. Archived from the orginal on 30.06.2015

Cuccuru, G., M. Orsini, A. Pinna, A. Sbardellati, N. Soranzo, A. Travaglione, P. Uva, G. Zanetti, and G. Fotia. 2014. Orione, a web-based framework for NGS analysis in microbiology. *Bioinformatics*. 30 (13): 1928-29.

Diaz, D.N., L. Krause, A. Goesmann, K. Niehaus, and T.W. Nattkemper. 2009. TACOA: taxonomic classification of environmental genomic fragments using a kernelized nearest neighbor approach. BMC *Bioinformatics*. 10: 56.

Drmanac, R., A.B. Sparks, M.J. Callow, A.L. Halpern, N. L. Burns, and B.G. Kermani, et al. 2010. Human genome sequencing using unchained base reads in self-assembling DNA nanoarrays. *Science* 327 (5961): 78–81.

Ewing, B., and P. Green. 1998. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res*. 8: 186–94.

Finn, R.D., J. Mistry, J. Tate, P. Coggill, A. Heger, J.E. Pollington, O.L. Gavin, A. Bateman. 2010. The Pfam protein families database. *Nucleic Acids Res*. 38: D211-D22.

George, I., B. Stenuit, and S. Agathos. 2010. Application of Metagenomics to Bioremediation. *Metagenomics: Theory, Methods and Applications*. Caister Academic Press.

Glass, E.M., J. Wilkening, A. Wilke, D. Antonopoulos, and F. Meyer. 2010. Using the metagenomics RAST server (MG-RAST) for analyzing shotgun metagenomes. Cold Spring Harb. Protoc., 1; pdb prot5368.

Gowda, M., H. Li, J. Alessi, F. Chen, R. Pratt, and G.L. Wang. 2006. Robust analysis of 5¢-transcript ends (5¢-RATE): a novel technique for transcriptome analysis and genome annotation. *Nucleic Acids Res*. 34 : e126.

Handelsman, J., M.R. Rondon, S.F. Brady, J. Clardy, and R.M. Goodman. 1998. Molecular biological access to the chemistry of unknown soil microbes: A new frontier for natural products. *Chem. Biol*. 5 (10): R245–R249.

Haque, M.M., T.S. Ghosh, D. Komanduri, and S.S. Mande. 2009. SOrt-ITEMS: Sequence orthology based approach for improved taxonomic estimation of metagenomic sequences. Bioinformatics. 25 (14): 1722-1730.

Harris. T.D., P.R. Buzby, H. Babcock, E. Beer, J. Bowers, and I. Braslavsky. et al. 2008. Single-molecule DNA sequencing of a viral genome. *Science* 320 (5872): 106–109.

Healy, F.G., R.M. Ray, H.C. Aldrich, A.C. Wilkie, L.O. Ingram, and K.T. Shanmugam. 1995. Direct isolation of functional genes encoding cellulases from the microbial consortia in a thermophilic, anaerobic digester maintained on lignocellulose. Appl. *Microbiol Biotechnol*. 43(4) : 667–74.

HeliScope Gene Sequencing / Genetic Analyzer System : Helicos BioSciences.

Hess, M., A. Sczyrba, R. Egan, T.W. Kim, H. Chokhawala, and G. Schroth. 2011. Metagenomic discovery of biomass degrading genes and genomes from cow rumen. *Science.* 331 (6016): 463-67.

Hoff, K.J., T. Lingner, P. Meinicke, and M. Tech. 2009. "Orphelia: predicting genes in metagenomic sequencing reads. *Nucleic Acids Res*. 37: W101-W105.

Hugenholz, P. 2002. Exploring prokaryotic diversity in the genomic era. *Genome Biology*. 3 (2): 1–8.

Huson, D.H., A.F. Auch, J. Qi, and S.C. Schuster. 2007. MEGAN analysis of metagenomic data. *Genome Res*. 17 (3): 377-86.

Hutchison, C.A. 2007. III, DNA sequencing: bench to bedside and beyond. *Nucleic Acids Res*. 35 : 6227–37.

Jaenicke. S., C. Ander, T. Bekel, R. Bisdorf, M. Dröge, K.H. Gartemann, S. Jünemann, O. Kaiser, L. Krause, F. Tille, M. Zakrzewski, A. Pühler, A. Schlüter, and A. Goesmann. 2011. Comparative and joint analysis of two metagenomic datasets from a biogas fermenter obtained by 454-pyrosequencing. *PLoS One*. 6(1): e14519.

Jay, E., R. Bambara, R. Padmanabhan, and R. Wu. 1974. DNA sequence analysis: a general, simple and rapid method for sequencing large oligodeoxyribonucleotide fragments by mapping. *Nucleic Acids Res*. 1 (3): 331–53.

Kanehisa, M., S. Goto, S. Kawashima, Y. Okuno, and M. Hattori. 2004. The KEGG resource for deciphering the genome. *Nucleic Acids Res*. 32 : D277-D80.

Kavita, K.S., C.L. Parsley, and M.R. Liles. 2010. Size does matter: application-driven approaches for soil metagenomics. *Soil. Biol. Biochem*. 42 (11): 1911–23.

Korbel, J.O., A.E. Urban, J.P. Affourtit, B. Godwin, F. Grubert, J.F. Simons, P.M. Kim, D. Palejev, N.J. Carriero, L. Du, B.E. Taillon, Z. Chen, A. Tanzer, A.C. Saunders, J. Chi, F. Yang, N.P. Carter, M.E. Hurles, S.M. Weissman, T.T. Harkins, M.B. Gerstein, M. Egholm, and M. Snyder. 2007. Paired-end mapping reveals extensive structural variation in the human genome. *Science* 318 : 420–26.

Krause, L., N.N. Diaz, A. Goesmann, S. Kelley, T.W. Nattkemper, F. Rohwer, R.A. Edwards, and J. Stoye. 2008. Phylogenetic classification of short environmental DNA fragments. *Nucleic Acids Res*. 36 (7): 2230-39.

Kristensen, D.M., A.R. Mushegian, V.V. Dolja, and E.V. Koonin. 2009. New dimensions of the virus world discovered through metagenomics. *Trends Microbiol*. 18 (1): 11–19.

Lai, B., R. Ding, Y. Li, L. Duan, H. Zhu. 2012. A de novo metagenomic assembly program for shotgun DNA reads. *Bioinformatics.* 28 (11) : 1455-62.

Lavery, T.J., B. Roudnew, J. Seymour, J.G. Mitchell, and T. Jeffries. 2012. High nutrient transport and cycling potential revealed in the microbial metagenome of australian sea lion (*Neophoca cinerea*) faeces. PLoS ONE. 7 (5): e36478.

Leung, H.C., S.M. Yiu, B. Yang, Y. Peng, Y. Wang, and Z. Liu. 2011. A robust and accurate binning algorithm for metagenomic sequences with arbitrary species abundance ratio. *Bioinformatics*. 27 (11): 1489-95.

Li., R., Y. Li, K. Kristiansen, and J. Wang. 2008. SOAP: short oligonucleotide alignment program. *Bioinformatics*. 24 (5): 713-14.

Liu, B., T. Gibbons, M. Ghodsi, and T. Treangen. 2011. Pop M: accurate and fast estimation of taxonomic profiles from metagenomic shotgun sequences. *BMC Genomics*. 12 (S 2): S4.

Luen-Luen, Li., S.R. McCorkle, S. Monchy, S. Taghavi, and D. van der Lelie. 2009. Bioprospecting metagenomes: glycosyl hydrolases for converting biomass. Biotechnol. *Biofuels* 2: 10.

Mardis, E.R. 2006. Anticipating the 1,000 dollar genome. *Genome Biol*. 7 : 112.

Margulies, M., M. Egholm, W.E. Altman, S. Attiya, and J.S. Bader, et al. 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*. 437: 376–80.

Markowitz, V.M., N.N. Ivanova, E. Szeto, K. Palaniappan, and K. Chu. 2008. IMG/M: a data management and analysis system for metagenomes. *Nucleic Acids Res*. 36 : D534-D38.

Maxam, A.M., and W. Gilbert. 1977. A new method for sequencing DNA. *Proc. Natl. Acad. Sci. U.S.A*. 74 (2): 560–64.

McHardy, A.C., H.G. Martin, A. Tsirigos, P. Hugenholtz, and I. Rigoutsos. 2007. Accurate phylogenetic classification of variable-length DNA fragments. *Nat. Methods*. 4 (1): 63-72.

Muller, J., D. Szklarczyk, P. Julien, I. Letunic, A. Roth, and M. Kuhn. 2010. eggNOG v2.0: extending the evolutionary genealogy of genes with enhanced non-supervised orthologous groups, species and functional annotations. *Nucleic Acids Res*. 38 : D190-D195.

Nakamura, K., T. Oshima, T. Morimoto, S. Ikeda, H. Yoshikawa, Y. Shiwa, S. Ishikawa, M.C. Linak, and A. Hirai, et al. 2011. Sequence-specific error profile of Illumina sequencers. *Nucleic Acids Res*. 39 (13): e90.

Namiki, T., T. Hachiya, H. Tanaka, and. Y. Sakakibara. 2012. MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads. *Nucleic Acids Res*. 40 (20) : e155.

Nederbragt , A.J. 2014. On the middle ground between open source and commercial software - the case of the Newbler program. *Genome Biol*. 15 (4): 113.

Noguchi, H., T. Taniguchi, and T. Itoh. 2008. MetaGeneAnnotator: detecting species specific patterns of ribosomal binding site for precise gene prediction in anonymous prokaryotic and phage genomes. *DNA Res*. 15 (6): 387-96.

Nyren, P., B. Pettersson, and M. Uhlen. 1993. Solid phase DNA minisequencing by an enzymatic luminometric inorganic pyrophosphate detection assay. *Anal. Biochem*. 208. 171–75.

Pace, N.R., E.F. Delong, and N.R. Pace. 1991. Analysis of a marine picoplankton community by 16S rRNA gene cloning and sequencing. *J. Bacteriol*. 173 (14): 4371–78.

Pace, N.R., D.A. Stahl, D.J. Lane, and G.J. Olsen. 1985. Analyzing natural microbial populations by rRNA sequences. *ASM News.* 51: 4–12.

Padmanabhan, R., and R. Wu. 1972. Nucleotide sequence analysis of DNA. IX. Use of oligonucleotides of defined sequence as primers in DNA sequence analysis. Biochem. Biophys. *Res. Commun*. 48 (5): 1295–302.

Padmanabhan, R., R. Wu, and E. Jay. 1974. Chemical synthesis of a primer and its use in the sequence analysis of the lysozyme gene of bacteriophage T4. *Proc. Nat. Aca*. Sci. 71 (6): 2510–14.

Patrick, D.S., and J. Handelsman. 2003. Biotechnological prospects from metagenomics. *Curr. Opin. Biotechnol*. 14 (3): 303–10.

Peng, Y., H.C. Leung., S.M. Yiu, F.Y. Chin. 2013. Meta-IDBA: a de Novo assembler for metagenomic data. *Bioinformatics*. 27 (13) : 194-101.

Porreca, G.J. 2010. Genome Sequencing on Nanoballs. *Nature Biotechnol*. 28 (1): 43–44.

Prober, J.M., G.L. Trainor, G.L. Dam, F.W. Hobbs, C.W. Robertson, and R.J. Zagursky, et al. 1987. A system for rapid DNA sequencing with fluorescent chain-terminating dideoxynucleotides. *Science*. 238 (4825): 336–341.

Raes, J., I. Letunic, T. Yamada, L. J. Jensen, and P. Bork. 2011. Toward molecular trait-based ecology through integration of biogeochemical, geographical and metagenomic data. *Mol. Sys. Bio*. 7: 473.

Rho, M., H. Tang, and Y. Ye. 2010. FragGeneScan: predicting genes in short and error-prone reads. *Nucleic Acids Res*. 38 (20) : e191.

Robertson, G., M. Hirst, M. Bainbridge, M. Bilenky, Y. Zhao, T. Zeng, G. Euskirchen, B. Bernier, R. Varhol, A. Delaney, N. Thiessen, O.L. Griffith, A. He, M. Marra, M. Snyder, and S. Jones. 2007. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat. Methods*. 4 : 651–657.

Ronaghi, M., S. Karamohamed, B. Pettersson, M. Uhlen, and P. Nyren. 1996. Real-time DNA sequencing using detection of pyrophosphate release. *Anal Biochem*. 242: 84–89.

Rusk, N. 2011. Torrents of sequence. *Nat. Meth*. 8 (1): 44–44.

Sanger, F., and A.R. Coulson. 1975. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J. Mol. Biol*. 94 (1975): 441–448.

Sanger, F., S. Nicklen, and A.R. Coulson. 1977. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U.S.A*. 74 (12): 5463–5467.

Schadt, E.E., S. Turner, and A. Kasarskis. 2010. A window into third-generation sequencing. *Hum. Mol. Gen*. 19 (R2): R227–40.

Selengut, J.D., D.H. Haft, T. Davidsen, A. Ganapathy, M. Gwinn-Giglio, and W.C. Nelson. 2007. TIGRFAMs and genome properties: tools for the assignment of molecular function and biological process in prokaryotic genomes. *Nucleic Acids Res*. 35 : D260-D264.

Shendure, J., G.J. Porreca, N.B. Reppas, X. Lin, J.P. McCutcheon, A.M. Rosenbaum, M.D. Wang, K. Zhang, R.D. Mitra, and G.M. Church. 2005. Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* 309 (5741): 1728–1732.

Soto-Jimenez, L.M., K. Estrada, and A. Sanchez-Flores. 2014. GARM: genome assembly, reconciliation and merging pipeline. *Curr Top Med Chem*. 14 (3) :418-424.

Stein, J.L., T.L. Marsh, K.Y. Wu, H. Shizuya, and E.F. DeLong. 1996. Characterization of uncultivated prokaryotes: isolation and analysis of a 40-kilobase-pair genome fragment from a planktonic marine archaeon. *J. Bacteriol*. 178 (3): 591–599.

Suen, G., J.J. Scott, F.O. Aylward, S.M. Adams, S.G. Tringe, A.A. Pinto-Tomás, and C.E. Foster, et al. 2010. An insect herbivore microbiome with high plant biomass-degrading capacity. *PLoS Genet*. 6(9) : e1001129.

Tatusov, R.L., N.D. Fedorova, J.D. Jackson, A.R. Jacobs, B. Kiryutin, E.V. Koonin, D.M. Krylov, R. Mazumder, S.L. Mekhedov, A.N. Nikolskaya, B.S. Rao, S. Smirnov, A.V. Sverdlov, S. Vasudevan, Y.I. Wolf, J.J. Yin, and D.A. Natale. 2003. The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*. 4 : 41.

Tawfik, D.S. and A.D. Griffiths. 1998. Man-made cell-like compartments for molecular evolution. *Nat. Biotechnol*. 16 : 652–656.

Thomas, T., D.M.Z. Rusch, P.Y. DeMaere Yung, M. Lewis, A. Halpern, K.B. Heidelberg, S.P.D. Egan Steinberg, and S. Kjelleberg. 2010. Functional genomic signatures of sponge bacteria reveal unique and shared features of symbiosis. *ISME J*. 4(12): 1557-1567.

Thompson, J.F., and K.E. Steinmann. 2010. Single molecule sequencing with a HeliScope genetic analysis system. *Current Protocols in Molecular Biology*. Chapter 7: Unit7.10.

Treangen, T.J., S. Koren, D.D. Sommer, B. Liu, I. Astrovskaya, B. Ondov, A.E. Darling, A.M. Phillippy, and M. Pop. 2013. MetAMOS: a modular and open source metagenomic assembly and analysis pipeline. *Genome Biol*. 14 (1): R2.

Tu, Q., Z. He, and J. Zhou. 2014. Strain/species identification in metagenomes using genome-specific markers. *Nucleic Acids Res*. 42(8): e67.

Tyson, G.W., J. Chapman, P. Hugenholtz, E.E. Allen, R.J. Ram, P.M. Richardson, V.V. Solovyev, E.M. Rubin, D.S. Rokhsar, and J.F. Banfield. 2004. Insights into community structure and metabolism by reconstruction of microbial genomes from the environment. Nature 428 (6978): 37–43.

United States1986. *Patent* 4,631,122.

Valouev, A., J. Ichikawa, T. Tonthat, J. Stuart, S. Ranade, and H. Peckham, et al. 2008. A high-resolution, nucleosome position map of *C. elegans* reveals a lack of universal sequence-dictated positioning. *Genome Res.* 18 (7): 1051–1063.

Venter, J.C., K. Remington, J.F. Heidelberg, A.L. Halpern, D. Rusch, and J.A. Eisen, et al. 2004. Environmental genome shotgun sequencing of the Sargasso Sea. *Science*. 304 (5667): 66-74.

Wang, X.V., N. Blades, J. Ding, R. Sultana, and G. Parmigiani. 2012. Estimation of sequencing error rates in short reads. *BMC Bioinformatics*. 13: 185.

Watson, J.D., and F.H. Crick. 1953. The structure of DNA. Cold. Spring. Harb. *Symp. Quant. Biol*. 18: 123–31.

Wong, D. 2010. Applications of Metagenomics for Industrial Bioproducts. *Metagenomics: Theory, Methods and Applications*. Caister Academic Press.

Wu, R., and E. Taylor. 1971. Nucleotide sequence analysis of DNA. II. Complete nucleotide sequence of the cohesive ends of bacteriophage lambda DNA. *J. Mol. Biol*. 57: 491–11.

Wu, R., C.D. Tu, and R. Padmanabhan. 1973. Nucleotide sequence analysis of DNA. XII. The chemical synthesis and sequence analysis of a dodecadeoxynucleotide which binds to the endolysin gene of bacteriophage lambda. Biochem. *Biophys. Res. Commun.* 55 (4): 1092–99.

Wu, Y.W., M. Rho, T.G. Doak, and Y. Ye. 2012. Stitching gene fragments with a network matching algorithm improves gene assembly for metagenomics. *Bioinformatics*. 28 (18) : 1363-1369.

Yok, N.G., and G.L. *Rosen*. 2011. Combining gene prediction methods to improve metagenomic gene annotation. *BMC Bioinformatics*. 12: 20.

Zerbino, D.R., and E. Birney. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res*. 18 (5): 821-29.

Zheng, H., and H. Wu. 2010. Short prokaryotic DNA fragment binning using a hierarchical classifier based on linear discriminant analysis and principal component analysis. *J. Bioinform Comput. Biol*. 8 (6): 995-1011.

Zimmer, C. (13 July 2010). How Microbes Defend and Define Us. *New York Times*. Retrieved 29 December 2011.